



National Institute of Environmental Health Sciences
Environmental Genome Project
NIEHS SNPs

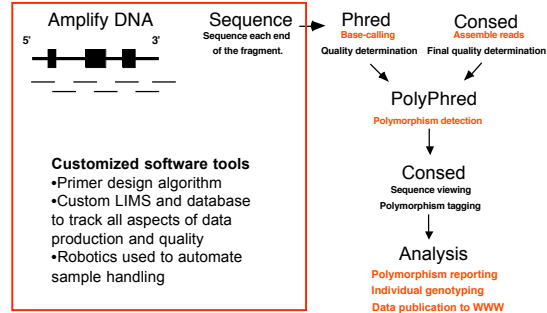
Search Site

NIEHS SNP Tutorial

Department of Genome Sciences
University of Washington

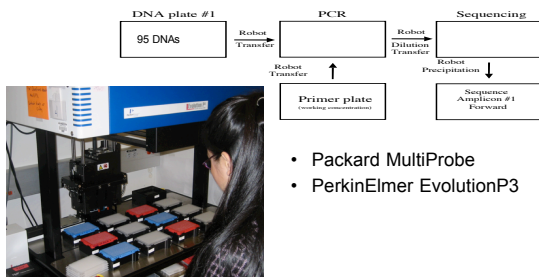
January 30-31, 2006

Sequencing production and data analysis pipeline



Robotics

Automated sample handling and plate setup

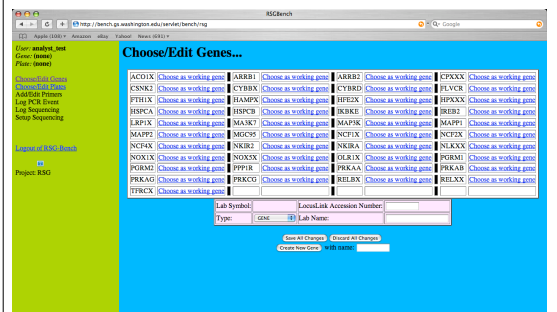


Data tracking using customized LIMS

- Tracks all aspects of sequencing production
 - Primer sequences
 - DNA samples
 - PCR quality
 - Inventories sequence chromatograms
 - Records read lengths, quality scores
 - Make sample tracking sheets for sequencers
 - All genotypes can be traced back to DNA sample

EGP Bench: Custom LIMS

Organized by gene



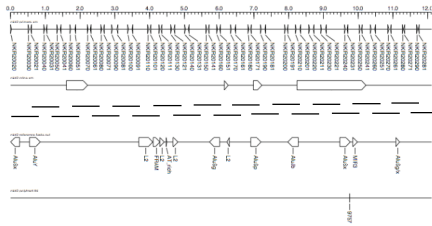
EGP Bench

Generates sample tracking form

Container Name	Plate ID	Description	Application	ContainerType	Owner	Operator	State/Labeling
SampleID:egp	egp_091805_06	Sequencing/Analysis	384-Well	ANALYST_TEST	ANALYST_TEST	Septa	
Well	Sample Name	Comment	Results Group	Instrument	Protocol 1	Analysis	Protocol 2
1234	PRAD0010001.0	137623.07	Sequencing/Analysis	Instrument	Protocol 3	Analysis	Protocol 4
1234	PRAD0010002.0	137623.07	Sequencing/Analysis	Instrument	Protocol 3	Analysis	Protocol 4
1234	PRAD0010003.0	137623.07	Sequencing/Analysis	Instrument	Protocol 3	Analysis	Protocol 4
1234	PRAD0010004.0	137623.07	Sequencing/Analysis	Instrument	Protocol 3	Analysis	Protocol 4
1234	PRAD0010005.0	137623.07	Sequencing/Analysis	Instrument	Protocol 3	Analysis	Protocol 4
1234	PRAD0010006.0	137623.07	Sequencing/Analysis	Instrument	Protocol 3	Analysis	Protocol 4
1234	PRAD0010007.0	137623.07	Sequencing/Analysis	Instrument	Protocol 3	Analysis	Protocol 4
1234	PRAD0010008.0	137623.07	Sequencing/Analysis	Instrument	Protocol 3	Analysis	Protocol 4
1234	PRAD0010009.0	137623.07	Sequencing/Analysis	Instrument	Protocol 3	Analysis	Protocol 4
1234	PRAD0010010.0	137623.07	Sequencing/Analysis	Instrument	Protocol 3	Analysis	Protocol 4
1234	PRAD0010011.0	137623.07	Sequencing/Analysis	Instrument	Protocol 3	Analysis	Protocol 4
1234	PRAD0010012.0	137623.07	Sequencing/Analysis	Instrument	Protocol 3	Analysis	Protocol 4
1234	PRAD0010013.0	137623.07	Sequencing/Analysis	Instrument	Protocol 3	Analysis	Protocol 4

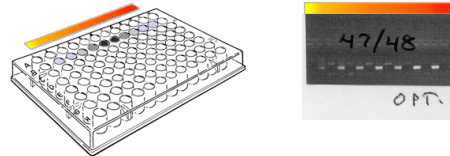
Re-sequencing pipeline

- Gene design- automated primer picking software
 - Exons, 2 kb upstream of first exon, 2 kb downstream of last exon
 - Genes larger than 30 kb have 10% of introns scanned
- Prior to amplification and re-sequencing, problematic GC-rich regions, alu repeats, polynucleotide tracts, and pseudogenes identified



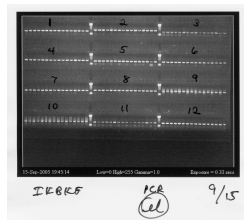
Re-sequencing pipeline

- PCR conditions optimized for each amplicon
- Failed optimization reactions repeated, primers redesigned upon second failure



Re-sequencing pipeline

- PCR quality spot checked prior to sequencing
- Failed PCR reactions repeated
- Refractory amplicons subjected to:
 - 10% DMSO
 - Alternate thermocycling parameters
 - Primer redesign

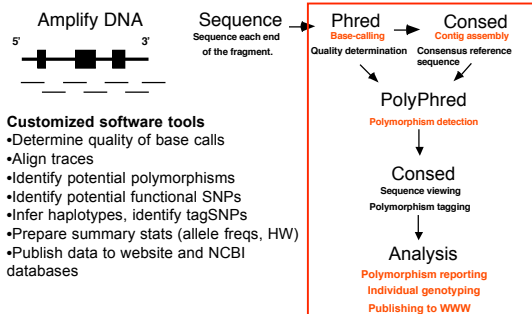


Re-sequencing pipeline

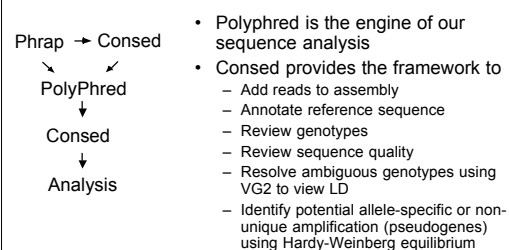


- Standard ABI BDT chemistry
 - Optimized for reaction volume and dilution
- Universal primer sequences standardize sequencing reaction conditions
- ABI 3730 capillary electrophoresis automated sequencers

Data analysis



Sequence Analysis



Extensive Quality Control Protocols and Checkpoints Built into the System

3 – PCR amplification

- Verification of PCR amplification and sizing
- Entry of PCR conditions and PCR results LIMS – linked to specific primers
- Robotic transfer of all DNA samples into pre-made, quality controlled PCR plates

4 - DNA sequencing

- Robotic transfer of all diluted PCR amplicons into pre-made, quality controlled sequencing plates
- Entry of sequencing reaction data into LIMS – linked to specific PCR amplicons and PCR events
- Generation of virtual barcode for each sequencing sample
- Automated generation of sequencing sample sheet (with virtual barcode)
- Daily sequencing reports automatically generated and emailed to laboratory technicians

Extensive Quality Control Protocols and Checkpoints Built into the System

5 - Gene assembly and polymorphism analysis

- Automated entry of sample chromatogram data in LIMS – linked to virtual barcode
- Automated entry of sample chromatogram QC data – Phred quality and read lengths
- Confirmation of orientation and location of sequence data on reference sequence during assembly
- Review of all tagged SNPs by data analyst to confirm quality
- Confirmation of all genotypes using double-stranded data
- Automated entry of polymorphism location and sample genotypes into LIMS

6 - Final data processing

- Confirmation of Hardy-Weinberg equilibrium for all sites (proportion of expected genotypes per site which can reveal problems stemming from allele-specific PCR amplification).

Data publishing

- Text files published to NIEHS SNPs web site and NCBI databases
 - SNP summary data
 - Genotypes
 - Final reference sequence
- Graphical data summaries with GeneSNPs and Visual Genotype images

NIEHS SNPs website

**National Institute of Environmental Health Sciences
Environmental Genome Project
NIEHS SNPs**

Welcome to the NIEHS SNPs Program

Introduction






The NIEHS Environmental Genome Project is a multi-disciplinary, collaborative effort focused on examining the relationships between environmental exposures, inter-individual sequence variation in human genes and disease risk in U.S. populations. The NIEHS SNPs Program at the University of Washington is targeted on the systematic identification and genotyping of single nucleotide polymorphisms (SNPs) in environmental response genes. The first phase of the effort is focused on finding common sequence variation (SNPs) in human genes involved in DNA repair and cell cycle pathways (see links under Gene Targets in the navigation menu on the left). Ultimately, the project will provide dense genetic maps of human genes that can be applied in evaluating human disease risk with environmental exposures.

GeneSNPs Database

NIEHS SNPs are available in the [GeneSNPs](#) database as well as the national database resource, [dbSNP](#). GeneSNPs provides a gene-centric map of the genome structure, coding sequences, and identified allelic variation in genes being targeted for a role in

Gene	Latest Update by Published Genes Table
CD8	December 26, 2005
FANCD2	December 27, 2005
MAK165	December 16, 2005
TM64	December 12, 2005
ALDH1A2	December 15, 2005
NEL3	December 1, 2005
ANKK10	November 30, 2005
CD40R2	November 30, 2005
DUSP1	November 30, 2005
IGFBP3	November 30, 2005
MEF	November 30, 2005
RETN	November 30, 2005
BCRA	November 30, 2005
TSP1R11	November 30, 2005
CYC1	November 22, 2005
CYP2B6	November 22, 2005
MT3	November 22, 2005
PLCD1	November 22, 2005
PRDX1	November 22, 2005
PRDX3	November 22, 2005
MAPK11	November 4, 2005
SHK11	November 4, 2005
PKCMB	November 2, 2005
TSUK2B	October 2, 2005

Data formats published to web facilitate association studies

LIMS-SPECIFIC LINKS				
Entire Gene	Golden Path (USBC Genome Browser)	Golden Path (with NIEHS SNPs Tracks)	Pub Med	
Download a zip file of all data for this gene			Example Population Descriptions	
 Mapping Data	cSNPs cDNA	Color FASTA SNP Context	PCR Primers (FASTA) Genbank	
 Genotyping Data	Visual Genotype Individual Genotypes	SNP Alleles SNP Allele Frequency	SNP Hardy-Weinberg	
 Haplotyping Data	PHASE Output Visual Haplotype	Phased Individual Haplotypes	Sorted by Frequency	
 Linkage Data	LD Select (Tag SNPs) African Descent	European Descent Hispanic Descent	Asian Descent	
 Predictive Analyses	Nonsynonymous cSNP Analysis			

Summary

Amplicons designed to tile across gene region using Tm- matched PCR primers

Amplicons sequenced using standard ABI BDT chemistry

Amplicon sequences assembled into contigs, annotated and reviewed using Consed

Polyphred 5.0 identifies potential SNPs, human reviewed

Custom LIMS tracks all aspects of data production and analysis

Rapid publishing of data files to web and national databases